The Development of Idiom Knowledge Across the Lifespan

Sprenger, Simone A.; la Roi, Amélie; van Rij, Jacolien

Published in:
Frontiers in Communication

DOI:
10.3389/fcomm.2019.00029

Publication date:
2019

frontiers in Communication

ORIGINAL RESEARCH
published: 05 July 2019
doi: 10.3389/fcomm.2019.00029

The Development of Idiom Knowledge Across the Lifespan

Simone A. Sprenger[1]*, *Amélie la Roi*[1] and *Jacolien van Rij*[2]*

[1] Center for Language and Cognition, Faculty of Arts, University of Groningen, Groningen, Netherlands, [2] Department of Artificial Intelligence, Faculty of Science and Engineering, University of Groningen, Groningen, Netherlands

Knowledge of multi-word expressions, such as *break the ice*, is an important aspect of language proficiency that so far we have known surprisingly little about. For example, it is largely unknown how much variability there is between speakers with respect to the number of different items that they know, or what factors contribute to their acquisition. This lack of knowledge seriously limits the generalizability of experimental studies on the production and comprehension of multi-word expressions (usually idioms) and generally suggests that there still is a sizable unknown territory of language knowledge to explore. Here, we present the results of two familiarity ratings for a large sample of Dutch idioms and a large number of participants that varied in age between 12 and 86 years old. The data show considerable variation between participants and between idioms. Non-linear mixed-effects regression analyses revealed that the age of participants, but not their education, as well as the frequency and decomposability of the idioms influenced the familiarity scores. Our findings suggest that the knowledge of multiword expressions develops across the lifespan, is acquired from exposure, and—in participants younger than about 40 years of age—varies with item decomposability.

Keywords: idioms, multiword expressions, vocabulary, aging, decomposability, development, familiarity

OPEN ACCESS

Edited by:
Gonia Jarema,
Université de Montréal, Canada

Reviewed by:
Giorgio Arcara,
IRCCS Fondazione Ospedale San Camillo, Italy
Debra Titone,
McGill University, Canada

***Correspondence:**
Simone A. Sprenger
s.a.sprenger@rug.nl
Jacolien van Rij
j.c.van.rij@rug.nl

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *07 November 2018*
Accepted: *04 June 2019*
Published: *05 July 2019*

Citation:
Sprenger SA, la Roi A and van Rij J (2019) The Development of Idiom Knowledge Across the Lifespan. Front. Commun. 4:29. doi: 10.3389/fcomm.2019.00029

INTRODUCTION

In everyday language use, many concepts are expressed by multi-word expressions, such as *hit the road* (depart), *break the ice* (relieve social tension by means of a remark) or *how are you* (a formula exchanged when people meet). These expressions must be learned alongside the words and rules that enable us to generate new sentences and represent an important aspect of what Pawley and Syder (1983) referred to as *nativelike* language proficiency. Based on analyses of conversational data, they estimated the number of such expressions in English as *hundreds of thousands* and suggested that access to these prespecified expressions in long-term memory is a prerequisite for fluent speech. Yet, even though the importance of multi-word expressions has been recognized in psycholinguistics (as evidenced by numerous experimental studies on the acquisition, processing and production of idioms, which we shortly discuss below), our knowledge about these processing units is still very limited. That is, in contrast to our knowledge about single words, we do not know what factors constrain the multi-word vocabulary and the way in which it varies between speakers. Here, we therefore want to explore how speaker characteristics (age, education) and item characteristics (frequency, decomposability) conjointly affect the acquisition of the Dutch idiom vocabulary across the lifespan.

In an exploration of what he called *the boundaries of the lexicon* (and thus the theoretical scope of grammatical theories), Jackendoff (1995) argued that the large number of multiword expressions

that speakers of a language know and recognize—which he estimated at about the same size as the number of single words—must in fact be considered entries in the mental lexicon. He illustrated his position with the *wheel of fortune* corpus, which included about 600 compounds, idioms, names and clichés, all considered sufficiently familiar to native speakers to be included in a popular TV game show that required participants to guess these phrases with a few hints. Examples include *I cried my eyes out*, *a breath of fresh air* and *May the Force be with you*.

While the nature of the underlying representations of such well-known phrases is still a matter of debate in linguistics and psycholinguistics (e.g., Cacciari and Tabossi, 1988; Fillmore et al., 1988; Cutting and Bock, 1997; Jackendoff, 1997; Titone and Connine, 1999; Sprenger et al., 2006; Libben and Titone, 2008), most idiom researchers will agree that they need to be included in the mental lexicon. However, our knowledge about this part of the lexicon is still limited. That is, we do not know how many multi-word expressions a speaker can be expected to be familiar with, or what this knowledge depends on. Estimates in the literature (such as Pawley and Syder, 1983, *hundreds of thousands*) are often extrapolations from small samples of conversation. At the same time, collections of multiword expressions in dictionaries or analyses of large corpora (e.g., Moon, 1998) can only provide upper boundaries for the knowledge that a native speaker might acquire. Neither method can provide us with a reliable estimate of the multiword vocabulary, or the conditions that affect its size.

Psycholinguistic approaches to multiword expressions typically focus on idioms. Apart from the fact that they form relatively fixed combinations of words, their meanings are not a direct function of their constituent words, making them an interesting test case for theories of language comprehension and production. For example, depending on the context, the English phrase *to break the ice* either refers to relieving the tension in a social situation or to the actual process of crushing frozen water. However, given a context that fits better with the figurative interpretation, native speakers can easily retrieve the correct form and meaning from memory (in production and comprehension, respectively).

Experimental work that tries to uncover the representations and processes that are responsible for the fast and efficient production and comprehension of idioms depends on high-quality stimulus materials. There are two main criteria that play role in this context: first, the idioms must be representative for a larger collection of items (e.g., with respect to the relationship between form and meaning), and second they must reflect the subjects' knowledge. This second criterion is especially difficult to fulfill. Does every speaker of English know the idiom *to kick the bucket,* or is that knowledge mostly restricted to the subset of idiom researchers? What other well-known expressions are there, and where are these items located in the frequency distribution? In idiom studies, questions about specific items are often answered on an *ad-hoc* basis, with stimulus materials being rated for familiarity in the context of a specific study. The number of items in these studies rarely exceeds twenty (e.g., Bobrow and Bell, 1973: 5 items; Swinney and Cutler, 1979: 22 items; Cacciari and Tabossi, 1988: 20 items; Cutting and Bock, 1997: 36 items; Gibbs, 1991: 20 items; Sprenger et al., 2006: 16 items;), and it is unclear in how far those are representative for the category of idioms as a whole. This lack of knowledge is a fundamental problem for psycholinguistic research on idiom production and comprehension, as it limits the potential generalizability of our data.

For various languages, such as English (Titone and Connine, 1994; Libben and Titone, 2008; Bulkes and Tanner, 2017; Nordmann and Jambazova, 2017), French (Caillies, 2009; Bonin et al., 2013, 2017), German (Citron et al., 2016), Italian (Tabossi et al., 2011) and Chinese (Li et al., 2016), norms have been published with the aim to increase the reliability of stimulus material in psycholinguistic studies on idioms. These norms provide a number of interesting measures, such as familiarity, decomposability, predictability or emotional valence, for several hundreds of items per language. That is, for the average speaker of the language in question, these norms provide a best guess about how a specific item scores on the various dimensions, making it possible for researchers to select items from the corresponding distributions.

However, while clearly increasing the reliability and validity of idiom tasks, the use of norms is not without problems either. It is important to realize that there is no such thing as an average native speaker: they differ with respect to socio-economic backgrounds, education, personality, and age. Given the effect of such variables on the sizes of our vocabularies at large (Brysbaert et al., 2016), it is conceivable that there are considerable individual differences in the idiom vocabulary as well. For example, Brysbaert et al. (2016) showed that the single-word vocabulary expands rapidly during adolescence, but keeps growing steadily until old age, with an average increase of one word per two days. In other words, age has an important effect on vocabulary that exceeds well-beyond the initial stages of language acquisition and cognitive maturation. Yet, idiom norming studies traditionally do not take this factor into account. They usually average across age, often sample from a student population only (e.g., Li et al., 2016), and sometimes do not mention their participants' age at all (e.g., Bulkes and Tanner, 2017). Whether age affects the idiom vocabulary in a similar way as the single-word vocabulary is therefore an open question.

Here, we want to explore the contribution of age to the development of the idiom vocabulary in more detail. If age indeed played an important role in idiom acquisition, this would have important consequences for the design of experiments that are to reveal the psycholinguistic processes and representations involved in the production and comprehension of idioms. Apart from the need to calibrate idiom norms for age, an age effect on idiom knowledge would stress the role of individual differences on online idiom comprehension. Reports on such effects so far have been few, but fairly consistent. Cain et al. (2005), for example, studied the relationship between reading comprehension and idiom interpretation in 9-year olds and found that poor comprehenders were less able to use context when interpreting opaque, but not transparent (or rather, decomposable) idioms. Cacciari et al. (2007) compared slow and fast participants in a comprehension task and found that slow participants needed more perceptual input to identify an idiom

and to activate its meaning. Columbus et al. (2015) found effects of executive control capacity on reading times for metaphors, but not for idioms. In contrast, Cacciari et al. (2018) found that idiom comprehension was affected by individual differences in working memory capacity, inhibitory control, and crystallized verbal intelligence, as well as personality-related variables (State Anxiety and Openness to Experience). Taken together, these studies indicate that individual differences affect online idiom comprehension processes, and thus are likely to affect acquisition as well. However, none of the studies considered age as a separate factor.

How would we expect age to affect the idiom vocabulary? First, the pattern that was observed by Brysbaert et al. (2016) for the development of the single-word vocabulary may be further delayed by the late development of figurative competence (i.e., the age at which children are able to understand an idiom's figurative interpretation, at about 9 years of age; Levorato and Cacciari, 1992), as well as by the relatively abstract concepts that are expressed by many idioms. So far, there are only few empirical data to backup this assertion, as developmental research on idioms has mostly focused on figurative competence, rather than the age at which children acquire specific tokens (e.g., Nippold and Martin, 1989; Levorato and Cacciari, 1992; Nippold and Rudzinski, 1993; Nippold and Taylor, 1995; Nippold and Duthie, 2003; Hung and Nippold, 2014). For example, Nippold and Martin (1989) report an increase in the ability to interpret idioms from the age of 14–17. As their observations are based on only twenty items per subject, we cannot draw conclusions about the size of the subjects' idiom vocabularies.

Beyond the age of adolescence, there are likewise only few data points to sketch the acquisition curve. A study by Kuiper et al. (2009) shows a rise in idiom knowledge until the age of 50–60 years, followed by a slight drop-off in the 65+ cohort (ten subjects per cohort). A drawback of this study is that the observations (based on 20 items) are not backed up by inferential statistics, making it difficult to judge their reliability. However, the pattern has partly been confirmed by Escaip (2015). Replicating Kuiper et al.'s (2009) study in Spanish, English, and French, she found a significant positive correlation of age with idiom knowledge in all three languages. That is, the older the participants, the more idioms they knew (with ages ranging between 15 and 83). For English, but not for the other two languages, Escaip also found a significant decrease of knowledge for speakers of 65 years and older.

The second important factor that we want to explore here is idiom frequency. In contrast to age, which is a characteristic of the subjects, frequency is a characteristic of the item itself. Similar to single-word acquisition, it is conceivable that frequency can explain a large part of the variance between idioms. This is supported by the observation that, in the past decade, a considerable number of studies has been published that demonstrate an important role for frequency in the acquisition of multi-word sequences. For example, Bannard and Matthews (2008) showed that children as young as 2 years old are sensitive to the frequency with which specific word combinations occur in child-directed speech: when asked to repeat sequences of words such as a *drink of tea*, they make fewer errors and—by the age of 3—also respond faster to high frequent word combinations than to matched low-frequent combinations. Likewise, Arnon and Snider (2010) demonstrated that adults are sensitive to the frequency of compositional multi-word phrases like *don't have to worry*. Their subjects responded faster in a phrasal decision task when the phrases were more frequent. Similar facilitatory effects for high-frequent items have been observed for language production in adult speakers, both for literal and more idiomatic sequences (e.g., Tremblay and Tucker, 2011; Janssen and Barber, 2012; Arnon and Cohen Priva, 2013; Sprenger and van Rijn, 2013).

The third factor that we include here is idiom decomposability, which was defined as the extent to which the idiom word meanings are related to the figurative meaning of the expression (similar to, for example, Rommers et al., 2013). Similar to frequency, decomposability is a feature of the individual idiom that may affect the ease with which a specific item can be acquired. If an idiom is highly decomposable, knowledge about its individual words may help the learner to deduce the idiom's meaning and/or to remember the item more easily when he or she encounters it again, since the words themselves may act as memory cues. This may explain why the poor comprehenders in the study by Cain et al. (2005) did not have difficulties interpreting decomposable idioms, in contrast to opaque idioms. From studies on online idiom processing, we know that decomposability is a relevant factor. Processing advantages have been reported for decomposable idioms over non-decomposable idioms: for example, with respect to sentence verification latencies (Gibbs et al., 1989) and in a lexical decision task that used idioms as primes for target words that were related to the item's figurative meaning (Caillies and Butcher, 2007). However, the exact nature of the way in which decomposability modulates online processing is still disputed, as its effect is not always facilitatory. Titone and Libben (2014) found late inhibitory effects of decomposability in a cross-modal semantic priming task and Titone et al. (2019) observed late inhibitory effects of decomposability during idiom reading. Interestingly, Westbury and Titone (2011) found an interaction of decomposability with age: in a literality judgment task, older adults were relatively slower than younger adults to accept non-decomposable idioms with a literal meaning and made more errors.

In the present article, we want to study the effect of age as an easy to assess speaker characteristic on idiom familiarity and compare it to the effects of idiom frequency and decomposability. If idioms indeed have their own entries in the mental lexicon, the idiom familiarity curve should be highly similar to that for single-word vocabulary (across speakers and items). That is, it should be modulated by age and education, with an early phase of rapid expansion, followed by a long phase of moderate but steady increase, and possibly decrease (as in Kuiper et al., 2009). Per item, this effect should be modulated by frequency, as we can expect the probability of acquisition to be a function of exposure. It may also be affected by idiom decomposability, which is supposed to reflect the ease with which an item can be analyzed, encoded, and retrieved (Caillies and Butcher, 2007). To test these predictions, we collected familiarity ratings for

194 Dutch idioms in two online rating studies and assessed the corresponding corpus frequencies. In addition to the ratings, respondents provided information about their gender, age, and level of education.

THE IDIOM DATABASE

For the exploration of the effect of age on idiom familiarity (Study 1 and 2 presented below), we have composed a small database with Dutch idioms. The database is available in the supplementary materials[1] and contains 189 Dutch idioms with their meaning and associated frequency counts. For all idioms (and control items, as explained below) we additionally collected decomposability ratings in an online questionnaire.

Materials and Methods

Materials

Ninety-nine Dutch idioms with two nouns were collected for Study 1. They were not controlled for syntactic structure or position of the nouns, but often contained prepositional phrases. The number of nouns was controlled with respect to the item's usability in an unrelated behavioral experiment. In addition to the experimental items, four German idioms were literally translated to Dutch and included as control items. All items were presented in past tense and preceded by the temporal adverb "Toen": (at a time in the past), for example "Toen kwam de aap uit de mouw." (*Then the monkey came out of the sleeve*, which means that the true nature of a situation, the true character of a person, or a hidden motive was being revealed).

Ninety Dutch idioms with one noun were collected for Study 2. Again, syntactic structure or noun position were not controlled for. All items were presented in past tense and preceded by the temporal adverb "Toen" (at a time in the past), for example "Toen zette hij hem op straat." (*Then he put him on the street*, which means *then he laid him off*) In contrast with Study 1, no control items were included. Thus, all idioms were existing Dutch idioms.

Frequencies for the idioms and translated German idioms were obtained from the Lassy Large corpus (Van Noord et al., 2013), a 700-million-word corpus of Dutch texts with automatically assigned syntactic annotations that is combined of both spoken and written sub-corpora (including the Dutch *Wikipedia*). By searching for lemmas, rather than exact word matching, most idioms were detected: the counts ranged between 0 (4 items) and 4,688. Surprisingly, three of the five control also were found in the corpus, probably due to their similarity with other Dutch idioms (for example, the German idiom *Then he shot sparrows with cannons* is very similar to the Dutch idiom *Then he shot mosquitos with cannons*). Before analysis, the frequency counts were log-transformed. **Figure 1** shows the distribution of the log-transformed frequencies.

Participants

The decomposability questionnaire was advertised under students of the University of Groningen. We restricted the age range to 18–25 years old, to keep the decomposability ratings consistent with earlier studies (Rommers et al., 2013). The data consisted of 57 entries, but we excluded one participant who was not monolingual Dutch (a Frisian-Dutch bilingual), 21 participants who contributed less than ten ratings, and one participant whose age did not match the target age range. The clean data consisted of 34 participants in the age range 21–26 years old (mean 24.3 years old; 8 men) who contributed each 15–98 ratings (mean 89.9). Participants did not receive compensation for their participation.

Procedure

The questionnaire was implemented using the survey software Qualtrics (Qualtrics, Provo, UT). Participants could anonymously access the questionnaire with a link. At the start of the experiment, participants were informed on the goal of the survey and gave their consent that their participation was voluntary. Participants were asked to read idioms and to judge to what extent the meaning of the individual words was related to the figurative meaning of the expression as a whole (cf. Rommers et al., 2013). They had to click on one of five radio buttons, labeled from left to right as "1 (geen relatie tussen individuele woorden en figuurlijke betekenis)" (*no relation between the individual words and the figurative meaning*), "2," "3," "4," and "5 (sterke relatie tussen individuele woorden en figuurlijke betekenis)" (*strong relation between the individual words and the figurative meaning*), or on a sixth radio button labeled as "ik ben niet bekend met deze uitdrukking" (*I am not familiar with this idiom*). Three idioms were presented individually at the start of the questionnaire to serve as anchors for the range of the decomposability scale (*anchoring*), but later idioms were presented in a random order. The idioms were divided in two lists of each 100 items (including the anchors). Each participant saw only one of the two lists.

Analyses

The data were analyzed using Generalized Additive Mixed Models (Hastie and Tibshirani, 1990; Wood, 2017; GAMMs), a *non-linear* mixed-effects regression method. GAMMs do not assume a linear relationship between the dependent variable and a covariate, but the relationship is estimated using penalized regression splines. The method does not require the user to specify the shape of the regression line on beforehand, but it is estimated based on the data. Other reasons for choosing this non-linear regression method are that it allows to include tensor product interactions for estimating interactions between multiple non-linear covariates, and it allows to include non-linear random effects (see for introductions Wieling, 2018; van Rij et al., in press). The statistical analyses are performed in R version 3.4.4 (2018-03-15; R Core Team, 2018), using the package *mgcv* version 1.8-24 (Wood, 2017) implementing GAMMs, and the package *itsadug* 2.3 (van Rij et al., 2017) for evaluation and visualization of the statistical models.

[1]Supplementary Materials are available at https://git.lwp.rug.nl/p251653/development-idiom-knowledge

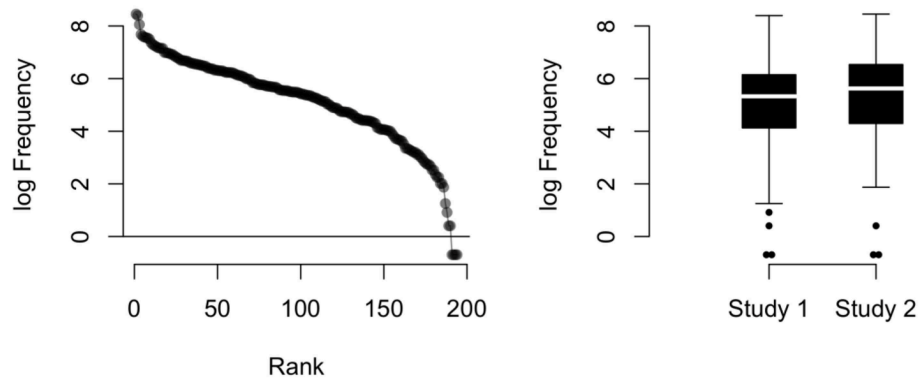

FIGURE 1 | Left: Frequency (log-transformed) of the 194 Dutch idioms and translated German idioms, ordered. **Right:** Comparison of the frequency distributions in Study 1 and Study 2.

Decomposability Ratings

From the 3,056 responses, 504 (16.5%) were of the category *I am not familiar with this idiom* (henceforth "unfamiliar" responses). These responses were excluded from the analysis. A logistic mixed-effects regression analysis revealed that the proportion of "unfamiliar" responses was significantly influenced by the idioms' frequencies [$\chi^2_{(2)} = 24.02$, $p < 0.001$]: the proportion of "unfamiliar" responses is larger for low-frequent idioms than for high-frequent idioms (see Supplementary Materials for the complete analysis).

All idioms were seen by at least thirteen participants. However, the number of actual decomposability ratings (i.e., when participants did *not* give an "unfamiliar" response) varied strongly between idioms, ranging from 2 to 34 (mean 13.1). **Figure 2** shows this variation in the number or ratings that was collected for each idiom: On the right end of the x-axis, there is one idiom that received a decomposability score from all 34 participants, because it was included as anchor. At the left end of the x-axis we find one of the translated German idioms with no decomposability ratings. All 13 participants who were presented with this idiom indicated that they were not familiar with it. We excluded this item from further analysis, accordingly.

We did not use the average rating per idiom as decomposability score, to avoid a potential subject bias influencing the decomposability scores for the idioms with a low number of ratings. Instead, we fitted a GAMM with random effects for participants and idioms to account for the participants' response biases and the variation between idioms. Random effects allow for partial pooling: the estimates for idioms that only have a few observations will pull toward the average (shrinkage); and the idiom estimates may be corrected for subject biases, as the subject mean is taken into account. From this statistical model we extracted an estimated decomposability score for each idiom (the script is available in the Supplementary Materials). To fit the ordered categorical nature of the decomposability ratings (5-point scale), we used the GAM ordered categorical family (Wood et al., 2016).

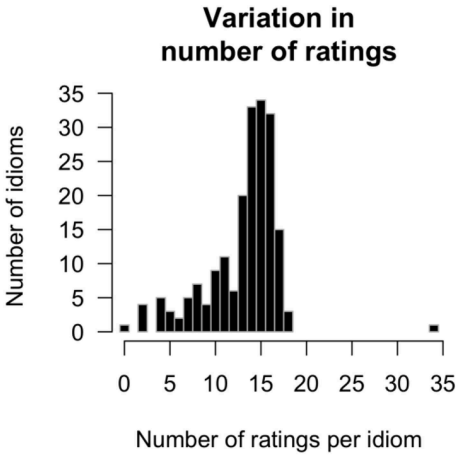

FIGURE 2 | Histogram of the number of decomposability ratings per idiom. The x-axis shows the number of decomposability ratings, and the y-axis the number of idioms that received that number of ratings.

Figure 3 (left panel) visualizes the difference between the mean rating scores (x-axis) and the estimated decomposability scores (y-axis). **Figure 3** (right panel) compares the estimated decomposability scores for Study 1 and Study 2.

Finally, we analyzed the effects of the idiom's frequency on the decomposability score. We used the GAM ordered categorical family (Wood et al., 2016) to fit the decomposability ratings (5-point scale). The log-transformed frequencies were included as non-linear main effect. In addition, by-Subject non-linear random smooths were included for Frequency and random intercepts for Idiom. However, the effect of Frequency was not significant [$F_{(1.001,2381.614)} = 2.69$; $p = 0.1$].

In the following sections we will use item frequencies and decomposability ratings as predictors for the familiarity ratings of Study 1 and Study 2.

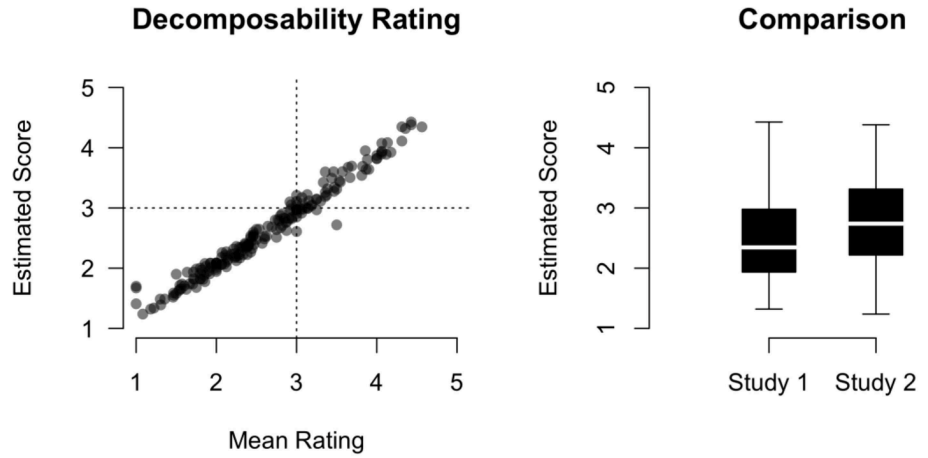

FIGURE 3 | Decomposability scores. **Left:** Mean rating scores (x-axis) vs. estimated rating scores with partial pooling (y-axis). The dashed lines mark the center of the scale. **Right:** Comparison of the estimated rating scores in Study 1 and Study 2.

STUDY 1: TWO-NOUN IDIOMS

In the first online questionnaire, we collected familiarity ratings for 104 Dutch idioms and control items with two nouns.

Materials and Methods

Participants

The questionnaire was advertised via social media (Facebook and Whatsapp) in the personal networks of the first and last author. The data consisted of 319 entries, but we excluded 25 participants who were not monolingual Dutch (13 of which were Frisian-Dutch bilingual). Subsequently, 37 participants were removed because the participants contributed less then ten ratings. The clean data consisted of 257 participants in the age range 12–86 years old (mean 37.7; 65 men) who contributed 96–104 ratings. Participants did not receive compensation for their participation.

Materials and Design

Ninety-nine Dutch idioms with two nouns were collected for this study. In addition, five German idioms were literally translated to Dutch and included as control items. The form of the materials *(Then the monkey came out of the sleeve)* was identical to the one described in section The idiom database.

All participants saw the same 104 idioms, but the order or presentation was randomized per participant. In addition to rating the idioms, participants were asked background questions about their gender, the year and month of birth, and their highest completed education (elementary school, high school, vocational education, or university).

Procedure

Participants could perform the questionnaire online on their computer, laptop, or tablet, or smartphone. The type of device was not registered. We have implemented the questionnaire using the survey software Qualtrics. Participants could anonymously access the questionnaire with a link. At the start of the experiment, participants were informed on the goal of the survey and they gave their consent that their participation was voluntary.

Participants were asked to read idioms and to judge whether their age peers would recognize this idiom when it would be used in a talk show. They had to click on one of five radio buttons, labeled from left to right as "1 (nog nooit gehoord)" (*never heard before*), "2," "3," "4," and "5 (heel bekend)" (*very well-known*). Three idioms were presented individually at the start of the questionnaire to serve as anchors for the range of the familiarity scale (*anchoring*), but later idioms were presented all at once in a long list in a random order to reduce the number of mouse clicks.

Results

Figure 4 shows the average rating per participant, plotted against their age (Left panel), the average rating per idiom and participant age (Center panel), and the average rating per age, collapsed over participants and idioms (Right panel). What immediately stands out from these plots is the variation between participants and between items. A closer look reveals that with younger ages the variation is larger than with older ages. Finally, the grand averages show us a clear increase in idiom familiarity over age, which continues in older ages. The ratings for each education level and the average age per education level are presented in **Table 1**.

The data were analyzed using GAMMs (Hastie and Tibshirani, 1990; Wood, 2017). We included *Education* and *Gender* as categorical predictors in the statistical model. Education is a three-level predictor describing the participant's education using the categories "University," "Vocational education," and "Other" (collapsing elementary school and high school). Further, we included the covariates *Age*, the participant's age in years, *Frequency*, the log-transformed frequency of the idiom, *Decomposability*, the estimated decomposability scores, and their interactions, and by-participant random smooths over Frequency and over Decomposability, and by-idiom random smooths over Age. These three random smooths account for variations between

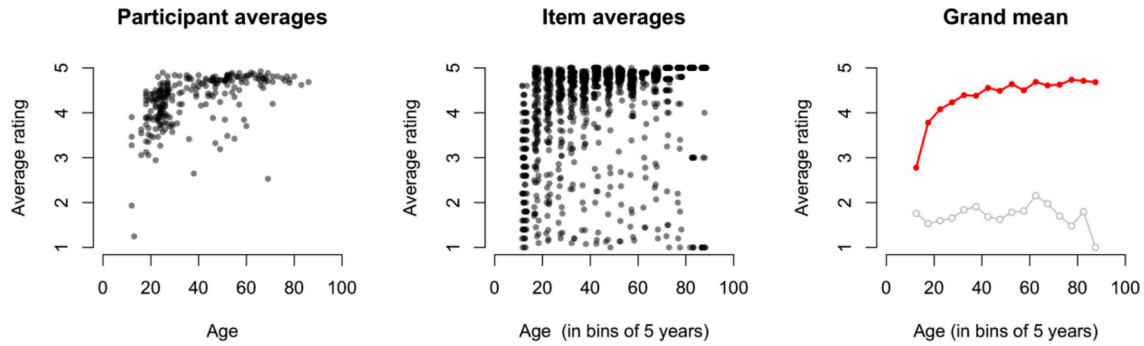

FIGURE 4 | Average familiarity ratings in study 1. **Left:** Average familiarity ratings per participant (y-axis) and the participant's age (x-axis). **Center:** Average familiarity ratings per idiom (y-axis) per age group (x-axis; in bins of 5 years, and jitter added for presentation purposes). **Right:** Grand average of familiarity ratings (y-axis) per age group (x-axis; in bins of 5 years). The solid red dots plot the grand average collapsed over all idioms, and the gray circles plot the grand average of the control items only.

TABLE 1 | Average ratings and average ages per education level in Study 1.

Education	*N*	Gender	Age		Rating
		Women	Mean	Median	
Elementary school	7	4	13.6	12	2.9
High school	39	31	31.8	21	4.2
Vocational education	35	27	50.7	51	4.6
University	176	130	37.4	29	4.3

participants and idioms, capturing random intercept differences and non-linear random deviations from the regression lines.

The dependent variable is the rating that participants provided for each idiom on a five-point rating scale (1 being unknown, and 5 being well-known). To account for the non-Gaussian nature of the dependent variable, the model was fitted using the GAM ordered categorical family, which implements regression for ordered categorical data (Wood et al., 2016). The smoothing parameter estimation method fREML was used, because the number of idioms and participants was too large for using ML. The method fREML allows for discretizing covariates and thereby decreasing the processing time enormously. A disadvantage of using fREML is that model comparisons are less reliable (see Wieling, 2018). Therefore, we did not only test a backward-fitting model comparison procedure using AIC and fREML, but also used the summary statistics and visual inspection of the model to determine the best-fitting model (cf. van Rij et al., 2019).

The manual backward-fitting model comparison procedure suggested that the non-linear three-way interaction between Age, Decomposability, and Frequency was not significantly contributing to the model [$\chi^2_{(4)} = 3.37$, $p > 0.1$; $\Delta\text{AIC} = -0.66$]. The summary statistics of the full model confirmed that the interaction surface was not significantly different from zero [$F_{(1.0,25792.4)} = 2.16$; $p > 0.1$). The non-linear two-way interaction between Decomposability and Frequency also did not show significance in the model comparison procedure [$\chi^2_{(3)} = 2.06$, $p > 0.1$; $\Delta\text{AIC} = -0.83$; summary statistics: $F_{(3.2,25792.4)} = 1.21$; $p > 0.1$]. The interaction between Age and Frequency was significantly contributing to the model [$\chi^2_{(3)} = 12.21$, $p < 0.001$; $\Delta\text{AIC} = -7.10$; summary statistics: $F_{(7.0,25792.4)} = 4.01$; $p < 0.001$]. The interaction between Age and Decomposability was found marginally significant [$\chi^2_{(3)} = 4.11$, $p = 0.42$; $\Delta\text{AIC} = 1.05$] in the model comparison procedure, but the summary statistics indicated that the interaction surface was different from zero [$F_{(1.0,25792.4)} = 7.48$; $p < 0.01$]. The categorical predictors Gender [$\chi^2_{(1)} = -0.44$, $p > 0.1$; $\Delta\text{AIC} = -0.02$] and Education [$\chi^2_{(2)} = 0.72$, $p > 0.1$; $\Delta\text{AIC} = 0.47$] did not contribute to the model, and were excluded from the model. The best-fitting model included the non-linear interactions between Age and Frequency and between Age and Decomposability and the non-linear main effects of Age, Frequency, and Decomposability. The best-fitting GAMM model: Rating ~ s(Age) + s(LogFreq) + s(Decomp) + ti(Age, LogFreq) + ti(Age, Decomp) + s(LogFreq, Subject, bs = 'fs', m = 1) + s(Decomp, Subject, bs = 'fs', m = 1) + s(Age, Sentence, bs = 'fs', m = 1), with the last three terms being non-linear random effects. In the best-fitting model, the main effects of Age [$F_{(3.2,25792.7)} = 31.30$; $p < 0.001$) and Frequency [$F_{(3.192,25792.736)} = 24.39$; $p < 0.001$] were significantly different from zero, but not the main effect of Decomposability [$F_{(1.1,25792.7)} = 3.56$; $p = 0.068$].

Figure 5 visualizes the estimates of the best-fitting GAMM by plotting the *fitted effects* (i.e., the sum of all model terms, which results in the model's estimate of the familiarity rating). The left panel shows the estimated effect of Age on the familiarity rating: the familiarity increases with age until around 60 years. Note that the values of the fitted effects are not directly comparable with the rating scale, because ordered categorical GAMMs use transformed values. The estimated cut-off points are added in red and these indicate how the transformed values relate to the response ratings. The Center panel shows the interaction between Age and Log Frequency in a contour plot, with on the z-axis the model's estimates for the familiarity ratings, again on the transformed scale. The interaction surface shows that for medium and high frequency values, the familiarity increases with age and is at ceiling for older participants.

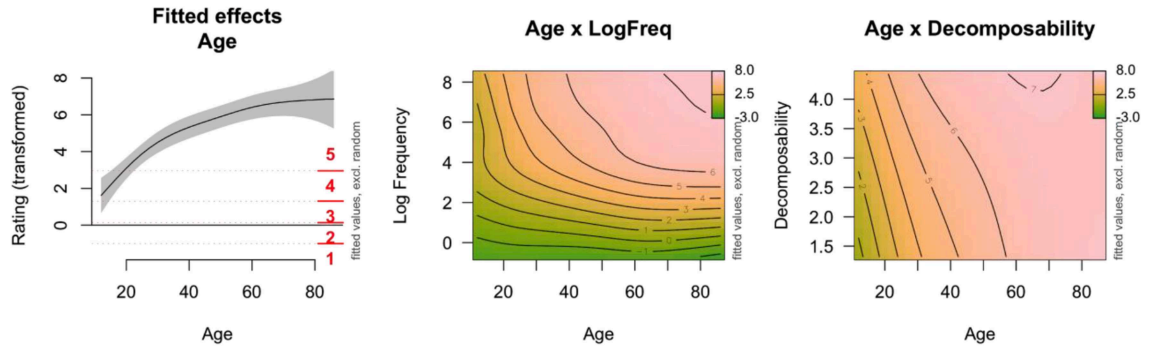

FIGURE 5 | Fitted effects of GAMM analysis of Study 1. Random effects are set to zero. **Left:** Estimated effect of Age (with pointwise 95% confidence intervals) for median values of LogFrequency (5.35) and Decomposability (2.35). Age is presented on x-axis, and the estimated familiarity rating on the y-axis. The red horizontal lines indicate the estimated cut-off points that relate the estimated values to the five response options. **Center:** Non-linear interaction between Age (x-axis) and Log Frequency (y-axis). The colors and contour lines indicate the value of the estimated ratings (z-axis). **Right:** Non-linear interaction between Age (x-axis) and Decomposability (y-axis). The colors and contour lines indicate the value of the estimated ratings (z-axis).

However, for the lowest frequency values, all age groups respond with a low familiarity value (i.e., the horizontal lines at the bottom). This is probably caused by the low frequency and control items, which also were rated as unfamiliar by the older participants. The Right panel visualizes the interaction between Decomposability and Age. Idioms with low decomposability scores are rated lower in familiarity than idioms with high decomposability scores. However, this decomposability effect is only found for younger and middle-aged adults, not for the older adults (> 60 years).

STUDY 2: ONE-NOUN IDIOMS

To verify whether the age effect also applies to other idioms and participants, we ran a second online questionnaire in which familiarity ratings for Dutch idioms were collected. This time the idioms had a different structure: instead of two nouns, the majority of these idioms contained one noun. The procedure of the experiment was exactly the same, only the participants and materials were different.

Materials and Methods

Participants

The questionnaire was advertised via social media (Facebook and Whatsapp) in the personal networks of the second author. The data consisted of 173 entries, but we excluded 56 participants who were not monolingual Dutch (48 of which were Frisian-Dutch bilingual). Subsequently, 12 entries were removed because the participants contributed less than ten ratings. The clean data consisted of data from 105 participants in the age range 19–76 years old (mean 42.9; 20 men) who contributed 15–90 ratings. Participants did not receive compensation for their participation.

Materials and Design

Ninety Dutch idioms with one noun were collected for this study. All items were presented in past tense and preceded by the temporal adverb "Toen" (at a time in the past), for example "Toen zette hij hem op straat." (*Then he put him on the street*, then he laid him off) In contrast to Study 1, no control items were included. Thus, all idioms were existing Dutch idioms.

All participants saw the same 90 idioms, but the order of presentation was randomized per participant. In addition to rating the idioms, participants were asked background questions about their gender, the year and month of birth, and their highest completed education (elementary school, high school, vocational education, or university).

Results

Figure 6 shows the average familiarity rating over age for participants (left panel), for idioms (center panel), and the grand average, collapsed over participants and idioms (right panel). Again, we see a large variation between participants and between items, maybe even more than in the data of Study 1 (**Figure 4**). The right panel shows the average rating per age, collapsed over participants and idioms. The plot shows an increase in average rating with age until the age of 55, after which the average ratings decrease again. This decrease was not visible in the averages of the data from Study 1. The ratings for each education level and the average age per education level are presented in **Table 2**.

The data of Study 2 were analyzed in the same way as Study 1, using Generalized Additive Mixed Models (Hastie and Tibshirani, 1990; Wood, 2017; GAMMs). We included *Education* and *Gender* as categorical predictors in the statistical model. Education is a three-level predictor describing the participant's education using the categories "University," "Vocational education," and "Other" (collapsing elementary school and high school). Further, we included the covariates *Age*, the participant's age in years, *Frequency*, the log-transformed frequency of the idiom, *Decomposability*, the decomposability scores, and we included as random effects by-participant random smooths over Frequency, by-participant random smooths over Decomposability, and by-idiom random smooths over Age.

The dependent variable is the ratings that participants provided for each idiom on a five-point rating scale (1 being

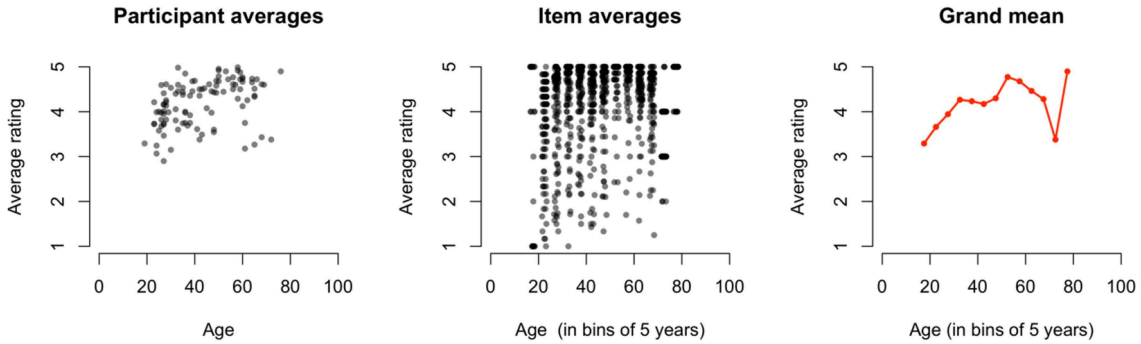

FIGURE 6 | Average familiarity ratings in study 2. **Left:** Average familiarity ratings per participant (y-axis) in relation to and the participant's age (x-axis). **Center:** Average familiarity ratings per idiom (y-axis) per age group (x-axis; in bins of 5 years, and jitter added for presentation purposes). **Right:** Grand average of familiarity ratings (y-axis) per age group (x-axis; in bins of 5 years).

TABLE 2 | Average ratings and average ages per education level in Study 2.

Education	*N*	Gender	Age		Rating
		Women	Mean	Median	
High school	7	7	52.0	64	4.1
Vocational education	28	24	45.3	47	4.1
University	70	54	41.1	37.5	4.2

unknown, and 5 being well-known). To account for the non-Gaussian nature of the dependent variable, the model was fitted using the GAMM ordered categorical family, which implements regression for ordered categorical data. As before, the smoothing parameter estimation method fREML was used, because the number of idioms and participants was too large for using ML.

The manual backward-fitting model comparison procedure suggested that the non-linear three-way interaction between Age, Decomposability, and Frequency was not significantly contributing to the model [$\chi^2_{(4)} = 3.67$, $p > 0.1$; ΔAIC $= -0.44$; summary statistics: $F_{(6.4,8111.0)} = 1.51$; $p > 0.1$]. The non-linear two-way interaction between Decomposability and Frequency also did not show significance in the model comparison procedure [$\chi^2_{(3)} = 3.11$, $p > 0.1$; ΔAIC $= 0.19$; summary statistics: $F_{(1.0,8111.0)} = 2.71$; $p = 0.1$]. The interaction between Age and Frequency was significantly contributing to the model [$\chi^2_{(3)} = 5.23$, $p = 0.015$; ΔAIC $= 2.69$; summary statistics: $F_{(5.9,8111.0)} = 2.41$; $p = 0.018$], and also the interaction between Age and Decomposability [$\chi^2_{(3)} = 6.25$, $p = 0.006$; ΔAIC $= -3.15$; summary statistics: $F_{(2.07,8111.03)} = 6.85$; $p < 0.001$]. The categorical predictors Gender [$\chi^2_{(1)} = 1.11$, $p > 0.1$; ΔAIC $= 0.14$] and Education [$\chi^2_{(2)} = 1.30$, $p > 0.1$; ΔAIC $= -0.35$] did not contribute to the model, and were excluded from the model. The best-fitting model included the non-linear interactions between Age and Frequency and between Age and Decomposability and the non-linear main effects of Age, Frequency, and Decomposability. As a result, we ended with the same specification for the best-fitting GAMM model as in the analysis of Study 1: Rating $\sim$ s(Age) + s(LogFreq) + s(Decomp) + ti(Age, LogFreq) + ti(Age, Decomp) + s(LogFreq, Subject, bs ='fs', m = 1) + s(Decomp, Subject, bs = 'fs', m = 1) + s(Age, Sentence, bs = 'fs', m = 1), with the last three terms being non-linear random effects. The main effects of Age [$F_{(3.1,8106.6)} = 6.62$; $p < 0.001$] and Frequency [$F_{(2.5,8106.6)} = 17.3$; $p < 0.001$] were significantly different from zero, but not the main effect of Decomposability [$F_{(1.0,8106.6)} = 0.09$; $p > 0.1$].

Figure 7 illustrates the fitted effects estimates of the GAMM analysis of the familiarity ratings of Study 2. The main effects regression line for Age indicates that the ratings increase with age until age 55, and decrease a little for the oldest participants. The oldest participants also show largest uncertainty around the estimates, because there are not many participants around 70. The center panel of **Figure 7** shows the interaction between Age and Frequency: idioms with a lower frequency result in lower familiarity ratings than idioms with a higher frequency, but this effect is stronger for young participants. The right panel of **Figure 7** shows the interaction between Age and Decomposability. The plot suggests that the decomposability scores influence the familiarity ratings of younger participants (< 40 years old), but not of older participants.

COMPARISON OF STUDY 1 AND STUDY 2

The findings reported both in Study 1 and Study 2 suggest that the familiarity of idioms increases with age, idiom frequency, and decomposability score. To test whether the trends for age, frequency, and decomposability are the same in the two experiments, we compared the estimated effects of the best-fitting models. As we used ordered categorical GAMMs, we cannot compare the model estimates directly. Ordered categorical GAMMs model the effects on a continuous scale and estimate the cut-off points that define the boundaries between the categories on the rating scale. These cut-off points are different for the analysis of Study 1 (−1, 0.14, 1.31, 2.96) and Study 2 (−1, 0.06, 1.13, 2.47). Instead, we can extract from the model the probability of the ordered categorical variable being of the corresponding category, and compare these probabilities (Wood et al., 2016). **Figure 8** illustrates the relation between the fitted estimates over

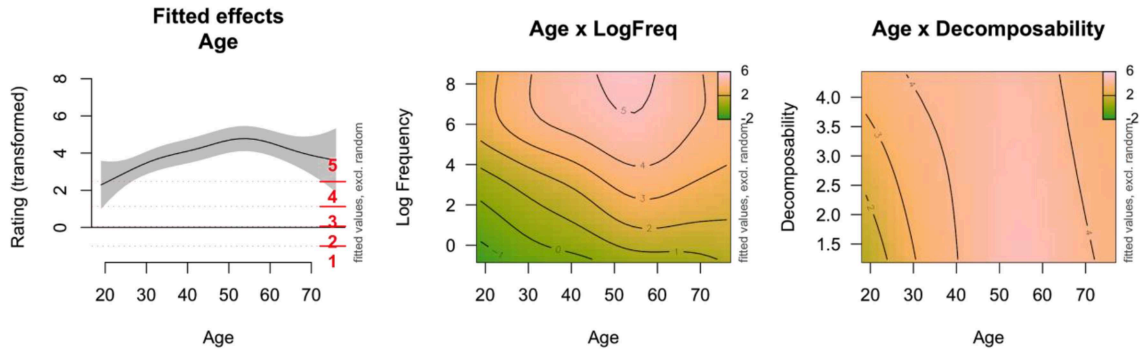

FIGURE 7 | Fitted effects of GAMM analysis of Study 2. Random effects are set to zero. **Left:** Estimated effect of Age (with pointwise 95% confidence intervals) for median values of LogFrequency (5.35) and Decomposability (2.35). Age is presented on x-axis, and the estimated familiarity rating on the y-axis. The red horizontal lines indicate the estimated the cut-off points that relate the estimated values to the five response options. **Center:** Non-linear interaction between Age (x-axis) and Log Frequency (y-axis). The colors and contour lines indicate the value of the estimated ratings (z-axis). **Right:** Non-linear interaction between Age (x-axis) and Decomposability (y-axis). The colors and contour lines indicate the value of the estimated ratings (z-axis).

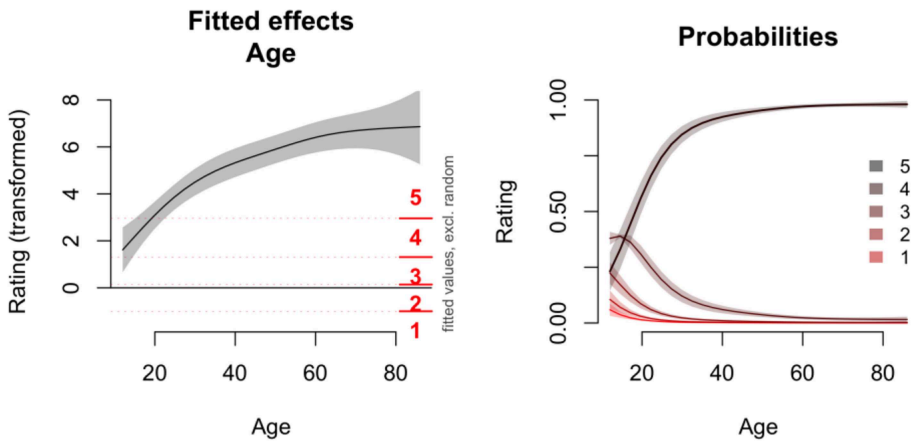

FIGURE 8 | The relation between fitted effects excluding random effects **(Left)**, and the probabilities for the dependent variable being in one of the five categories **(Right)**. The left panel includes the estimated cut-off points in red.

Age (summing over all predictors, including the intercept; Left panel), and the probabilities for the dependent variable being in one of the five categories, using the effect of Age in Study 1 (Right panel).

In the left panel of **Figure 9** we (visually) compared the effect of Age on the probabilities for the response variable being rating 4 or 5. To facilitate the comparison we did not include the other three ratings in the plot. The solid lines are the estimated probabilities based on the best-fitting model fitted on the data of Study 1, whereas the dashed red lines are the estimated probabilities based on the best-fitting model fitted on the data of Study 2. Over all ages, the probability of selecting 5 (very well-known) is higher in Study 1 than in Study 2, but the probability of selecting 4 is higher in Study 2 than in Study 1. Thus, irrespective of Age, the idioms in Study 2 are rated as less familiar than the idioms in Study 1. In addition, we see a clear decrease in selecting response option 5 for older adults (> 60 years) in Study 2, but not in Study 1. Important to mention is that these fitted effects are calculated for a median log-frequency (5.48) and a median decomposability score (2.50).

The center panels of **Figure 9** show the estimated probabilities of the response variable being rating 5 for the data of Study 1 (top) and Study 2 (bottom), and how this probability is influenced by Frequency and Age. The contour plot of Study 1 indicates that low frequency idioms (such as the control items, which were translated German idioms that do not exist in Dutch) are unfamiliar for all age groups, as indicated by the green color which is associated with low probabilities and the horizontal contour lines (i.e., no changes in Age, only in frequency). The high frequency idioms on the other hand are rated as being highly familiar by participants older than 30, as indicated by the pink color which marks a probability of 1. These same high frequency idioms show a sharp increase over age in the probability of being rated as 5 for participants under 30 years, as indicated by the vertical contour lines (i.e., no change in frequency, but in age). The contour plot of Study 2 roughly shows similar patterns (increase in probability with Age and Frequency), but

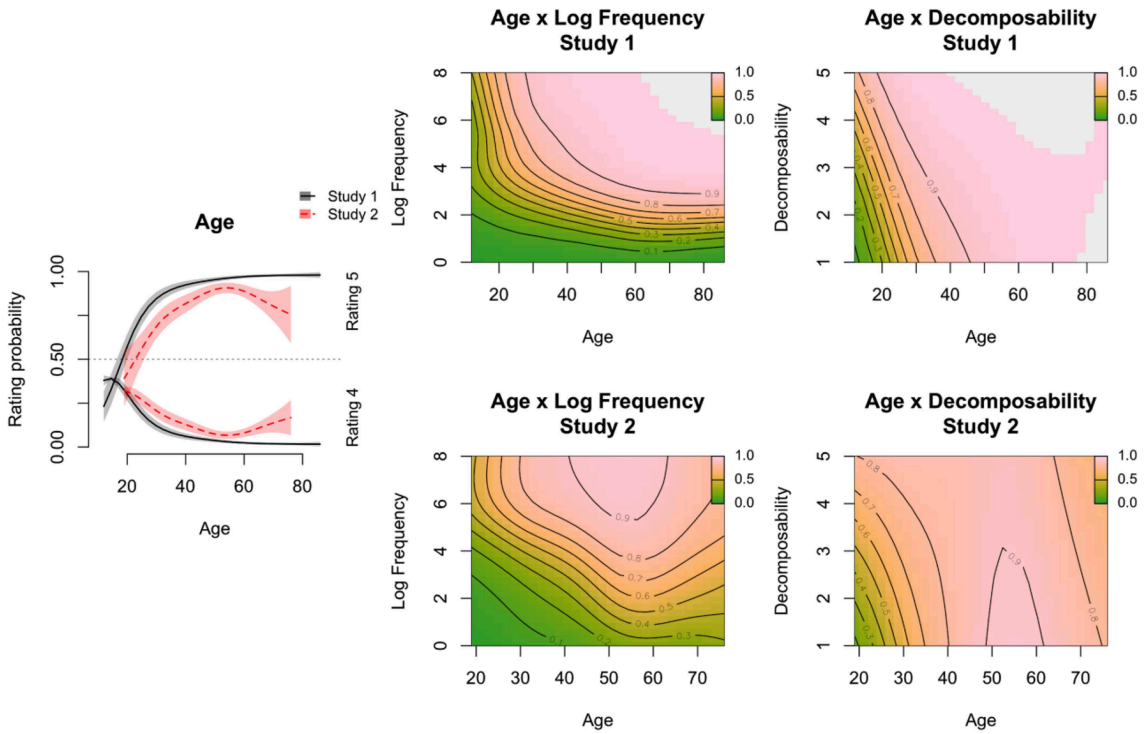

FIGURE 9 | Probabilities of the response variable being rating 5 (highly familiar). The probabilities for Studies 1 and 2 are derived from two different statistical models. Random effects are set to 0. **Left:** Comparing the probabilities of rating 4 and 5 (y-axis) over Age (x-axis) for Studies 1 and 2, with median values for Log Frequency (5.48) and Decomposability (2.50). **Center:** Effects of Age (x-axis) and Log Frequency (y-axis) on the probability of rating 5 (z-axis) for Study 1 (top) and for Study 2 (bottom) with Decomposability set to a median value (2.50). **Right:** Effects of Age (x-axis) and Decomposability (y-axis) on the probability of rating 5 (z-axis) for Study 1 (top) and for Study 2 (bottom) with Log Frequency set to a median value (5.48). Note that the age range is smaller for Study 2 than for Study 1.

the pattern looks more variable. One of the causes of the more variable pattern in Study 2 may be that no control items had been included.

The right panels of **Figure 9** shows the estimated probabilities of the response variable being rating 5 for the data of Study 1 (top) and Study 2 (bottom), and how this probability is influenced by Decomposability and Age. The two contour plots show a very similar pattern: idioms that are perceived as highly decomposable, have a higher probability of being rated with a 5 (highly familiar) by young and old participants. Idioms that are perceived as less decomposable, have a lower probability of being rated with a 5 by younger participants ($<$ 40 years old). On the basis of visual inspection, it seems that the two studies show a stronger difference in the interaction between Age and Frequency than in the effect of Decomposability.

DISCUSSION

We explored the variability in idiom knowledge in a large sample of native speakers of Dutch, divided across two separate idiom familiarity studies. Based on findings for single-word vocabulary, and driven by the assumption that idioms (and other multi-word expressions) can be considered entries in the mental lexicon, we expected to find a familiarity curve that shows an early phase of rapid expansion, followed by a long phase of moderate but steady increase across the lifespan. This pattern has been confirmed in both studies. The grand averages show a clear increase in idiom familiarity over age that proceeds until at least 55 years of age.

We also observed a predicted delay in the rise of the idiom vocabulary, compared to single-word acquisition, by about 10 years, as the steep increase in idiom acquisition levels off after 30 years.

The second main factor of interest was idiom frequency. As with single words, the simple rationale is that the more often speakers come across a specific item, the higher the probability of long-term retention. We therefore expected higher ratings for high-frequent idioms. This pattern has been confirmed in both studies, with frequency significantly impacting on the probability of a high familiarity score. While the low frequency items consistently score low in all age groups, the high frequency idioms are rated as being highly familiar by participants older than 30 and show a sharp familiarity increase over age for participants younger than 30 years. Based on Brysbaert et al.'s (2016) findings for the single-word vocabulary, we further expected to find evidence for an effect of education. However, no such effect was found.

As a third factor, we included independent ratings of idiom decomposability, as they might inform us about the way in which item characteristics affect the ease with which idioms are acquired across age. We indeed find that idioms with low decomposability scores are rated as less familiar than items with high decomposability scores, in both studies. However, this

effect seems restricted to the younger raters (< 40 years old). This suggests that the degree to which the individual words are perceived to contribute to the meaning of the idiom as a whole affects the item's learning trajectory. Decomposable idioms may be more easily acquired than non-decomposable idioms, which could explain why Cain et al. (2005) found that children with poor reading comprehension skills had difficulties interpreting non-decomposable, but not decomposable idioms. Whereas the meaning of decomposable idioms can be derived from the meanings of the idiom constituents, the meaning on non-decomposable idioms has to be learned explicitly. Yet, our findings also suggest that once the item has been acquired, the degree to which it is decomposable no longer affects its perceived familiarity. In this context, it is noteworthy that we deliberately limited the age range of the participants who provided the decomposability ratings (18–25 years), to avoid a possible confound of the ratings with age. An interaction of item decomposability and age has been reported for online processing (Westbury and Titone, 2011). In a follow-up on our study on idiom knowledge, it would be interesting to see to what degree offline decomposability judgements vary with age, as this might further affect the generalizability of many sets of idiom norms.

A comparison of the two studies (two-noun vs. one-noun idioms) revealed that the effects of Age and Frequency on the familiarity judgements in Studies 1 and 2 are roughly similar. The most important difference is a decrease in familiarity for the older participants (> 60 years) in Study 2, but not in Study 1. It is not clear what has caused this difference: The education levels of the older participants are very similar between studies (**Table 3**), and the predictors Education and Gender did not show an effect on the familiarity ratings in the statistical analyses. However, the number of older participants in Study 2 was much lower than in Study 1, and hence the variation between participants might have had a larger effect than in Study 1. In comparison, the difference in the effects of decomposability is relatively smaller.

Overall, the pattern of the idiom acquisition curve that we find in the two studies shows that—unlike what is often taken for granted in idiom processing studies—idiom knowledge varies widely between age groups. Especially young adolescents (students) cannot be expected to have developed a large idiom vocabulary yet. This implies that they constitute a relatively unreliable group for testing theories of idiom comprehension and production: they may or may not be familiar with the items, and their representations may be less stable than those of speakers above the age of 30.

TABLE 3 | Comparison of the levels of education of the older participants (> 60 years) in Study 1 and in Study 2.

	Study 1	Study 2
High school	6	4
Vocational education	9	4
University	21	11
Total	36	19

A possible explanation for the delay in idiom acquisition (in comparison to that of the single-word lexicon as described by Brysbaert et al., 2016) may be found both in the subject and item characteristics. First, idioms are figurative expressions and the ability to handle such expressions successfully only develops at around 9 years of age (Levorato and Cacciari, 1992). Second, idioms often tend to refer to relatively abstract and/or pragmatically complex concepts that may only be grasped well-beyond puberty. A third possibility is that what we observe in our data is in fact an indicator of language change. That is, the younger participants may simply not be familiar with the items because they are no longer being used and/or have been replaced by new idioms. Given the method of item selection (based on examples found in newspaper articles and conversations, but also idiom dictionaries), we do not find this explanation very likely, but we feel that it would be worth exploring in a future study. A methodological challenge will be however that any *new* idioms that would be expected to replace the old items in the vocabularies of the younger generations will first need to be identified.

At the other end of the distribution, old age, our findings are somewhat inconclusive: do elderly speakers experience problems in accessing items that they used to know before? Based on findings by Kuiper et al. (2009) and Escaip (2015) that showed mixed evidence for a late drop in idiom knowledge, we were especially interested in the category of 65+ participants. While our first study does not show any evidence for such a drop, the second study shows a slight decrease. Yet, the relatively few subjects in these categories and the large variability make it difficult to estimate the reliability of this effect. An additional explanation may be the influence of Frisian in this sample. Although we took care to remove all native speakers of Frisian, it is possible that the remaining participants are also predominantly located in Friesland and therefore come across different idioms in everyday life. For the older participants, this effect may be much stronger, as they can be expected to be less mobile and less exposed to mainstream (Dutch) media. In a future study, we therefore need to include information about the subject's geographical location, about the area in which they grew up, and the type of media that they consume. Ideally, this would be a megastudy comparable to that of Brysbaert et al. (2016), with a large number of items and a very large and diverse sample of Dutch speakers.

This study provides support for the hypothesis that idiom acquisition is similar to word acquisition, with increasing knowledge across the life span. However, idioms are different from words in that they are multiword expressions, and idioms are different from many other types of multiword expressions in that they have a figurative meaning. It would be interesting to compare the acquisition of idioms with the acquisition of other types of multiword expression to investigate how the ability to understand figurative expressions influences idiom acquisition. Is this a prerequisite for acquiring idioms, as is generally assumed? Or do children acquire high frequent idioms as words, without the ability to understand figurative language? One of the difficulties in investigating these questions is the variability between idioms. Other factors such as concreteness and imageability (both related to the transparency of the idiom)

could play a role in whether and how much idioms are being perceived as figurative language. The effects that we find for decomposability support this hypothesis.

Taken together, our findings stress the need for future work to address both item and subject characteristics that could potentially affect idiom acquisition in more detail. Our findings with respect to the effect of decomposability and its interaction with age suggest that this could be a worthwhile enterprise. With respect to other item characteristics, possible candidates are, for example, the above mentioned factors concreteness, transparency, and imageability, but also length, or animacy.

While the idea that idioms differ with respect to item characteristics, such as decomposability, was formulated early on in the idiom literature (e.g., Gibbs and Nayak, 1989), the focus on speaker characteristics is relatively new (see also section Introduction). Yet, the idea that successful idiom comprehension depends on individual differences in processing abilities seems relatively straightforward, as idiom comprehension is a complex skill which is only acquired late during acquisition. For example, Cacciari et al. (2018) found a clear relationship between online idiom comprehension and cognitive functions that might extend to idiom acquisition as well, to the extent that it reflects differences in fluid intelligence. On top of that, the personality traits that they found to affect online processing might come into play in acquisition, too. The use of figurative language and other multi-word expressions is an important stylistic device that may be very well-suited to express different types of personality. The factor Age, which has been in the focus of the present article, may thus not only represent a participant's linguistic experience, but also its interaction with age-dependent changes in cognitive control, long-term memory access, and personality. Future studies will need to distinguish these factors on a more fine-grained level.

Are all native speakers alike when it comes to idioms? We have shown that—similar to the single-word vocabulary—the idiom vocabulary differs widely across speakers, with age rather than education being the main factor driving these differences. *Are all idioms alike when it comes to the probability of being known by a native speaker?* In line with findings on online processing (e.g., Arnon and Snider, 2010) we have shown that idioms behave much like ordinary entries in the mental lexicon, in that they are sensitive to distributional information. The more frequent an idiom, the larger the probability that a native speaker is familiar with it. In addition, the probability with which an idiom is acquired is affected by the degree to which it is decomposable. Our findings can help increase the reliability and validity of idiom processing studies. More importantly, we think that they contribute to a clearer picture of the way in which the boundaries of the lexicon expand across the lifespan.

DATA AVAILABILITY

The datasets generated for this study can be found here: https://git.lwp.rug.nl/p251653/development-idiom-knowledge.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the Ethical Rules for Conducting Research with Human Participants, Research Ethics Committee Faculty of Arts (CETO), University of Groningen. The protocol for all age groups was approved by the Research Ethics Committee Faculty of Arts (CETO), University of Groningen (60761519). All subjects gave written informed consent in accordance with the Declaration of Helsinki. We did not obtain written consent from parents/legal guardians of participants under 16, because participants participated anonymously in the experiment by clicking a link that was posted on social media, the participants were free to stop whenever they wished without consequences (they were not payed or reimbursed for their participation), and the materials of the experiment (officially recognized Dutch idioms) gave no reason to assume that non-adult participants could suffer negative consequences from participating in this study.

AUTHOR CONTRIBUTIONS

SS and JvR: conceptualization and writing. SS and AlR: materials and data collection. JvR: statistical analyses.

FUNDING

This research was supported by grants from the Netherlands Organization for Scientific Research NWO (Veni grant no. 275-70-044, JvR; and Ph.D. in the Humanities grant no. 322-75-008, AlR).

ACKNOWLEDGMENTS

We thank Prof. Dr. Gertjan van Noord and Peter Kleiweg for their assistance with the search for idioms in the Lassy Large corpus.

REFERENCES

Arnon, I., and Cohen Priva, U. (2013). More than words: the effect of multi-word frequency and constituency on phonetic duration. *Lang. Speech* 56, 349–371. doi: 10.1177/0023830913484891

Arnon, I., and Snider, N. (2010). More than words: frequency effects for multi-word phrases. *J. Memory Lang.* 62, 67–82. doi: 10.1016/j.jml.2009.09.005

Bannard, C., and Matthews, D. (2008). Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations. *Psychol. Sci.* 19, 241–248. doi: 10.1111/j.1467-9280.2008.02075.x

Bobrow, S. A., and Bell, S. M. (1973). On catching on to idiomatic expressions. *Mem. Cogn.* 1, 343–346.

Bonin, P., Méot, A., Boucheix, J. M., and Bugaiska, A. (2017). Psycholinguistic norms for 320 fixed expressions (idioms and proverbs) in French. *Quart. J. Exp. Psychol.* 71:1057–69. doi: 10.1080/17470218.2017.1310269

Bonin, P., Méot, A., and Bugaiska, A. (2013). Norms and comprehension times for 305 French idiomatic expressions. *Behav. Res. Methods* 45, 1259–1271. doi: 10.3758/s13428-013-0331-4

Brysbaert, M., Stevens, M., Mandera, P., and Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word

definition, the degree of language input and the participant's age. *Front. Psychol.* 7:1116. doi: 10.3389/fpsyg.2016.01116

Bulkes, N. Z., and Tanner, D. (2017). "Going to town": large-scale norming and statistical analysis of 870 American English idioms. *Behav. Res. Methods* 49, 772–783. doi: 10.3758/s13428-016-0747-8

Cacciari, C., Corradini, P., and Ferlazzo, F. (2018). Cognitive and personality components underlying spoken idiom comprehension in context. An exploratory study. *Front. Psychol.* 9:659. doi: 10.3389/fpsyg.2018.00659

Cacciari, C., Padovani, R., and Corradini, P. (2007). Exploring the relationship between individuals' speed of processing and their comprehension of spoken idioms. *Eur. J. Cogn. Psychol.* 19, 417–445. doi: 10.1080/09541440600763705

Cacciari, C., and Tabossi, P. (1988). The comprehension of idioms. *J Memory Lang.* 27, 668–683. doi: 10.1016/0749-596X(88)90014-9

Caillies, S. (2009). Descriptions de 300 expressions idiomatiques: familiarité, connaissance de leur signification, plausibilité littérale,≪ décomposabilité ≫ et ≪ prédictibilité ≫. *L'année Psychol.* 109, 463–508. doi: 10.4074/S0003503309003054

Caillies, S., and Butcher, K. (2007). Processing of idiomatic expressions: evidence for a new hybrid view. *Metaphor. Symbol.* 22, 79–108. doi: 10.1080/10926480709336754

Cain, K., Oakhill, J., and Lemmon, K. (2005). The relation between children's reading comprehension level and their comprehension of idioms. *J. Exp. Child Psychol.* 90, 65–87. doi: 10.1016/j.jecp.2004.09.003

Citron, F. M., Cacciari, C., Kucharski, M., Beck, L., Conrad, M., and Jacobs, A. M. (2016). When emotions are expressed figuratively: psycholinguistic and Affective Norms of 619 Idioms for German (PANIG). *Behav. Res. Methods* 48, 91–111. doi: 10.3758/s13428-015-0581-4

Columbus, G., Sheikh, N. A., Côté-Lecaldare, M., Häuser, K., Baum, S. R., and Titone, D. (2015). Individual differences in executive control relate to metaphor processing: an eye movement study of sentence reading. *Front. Hum. Neurosci.* 8:1057. doi: 10.3389/fnhum.2014.01057

Cutting, J. C., and Bock, K. (1997). That's the way the cookie bounces: syntactic and semantic components of experimentally elicited idiom blendsß. *Memory Cogn.* 25, 57–71. doi: 10.3758/BF03197285

Escaip, V. (2015). *The Relationship of Phrase Head Word Frequency and Acquired Idioms: A Comparative Analysis of Spanish, English and French Verb Phrase Idioms*. Unpublished thesis, University of Canterbury, New Zealand. Retrieved from: http://hdl.handle.net/10092/11757

Fillmore, C. J., Kay, P., and O'connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language* 64, 501–538.

Gibbs, R. W. (1991). Semantic analyzability in children's understanding of idioms. *J. Speech Language Hearing Res.* 34, 613–620. doi: 10.1044/jshr.3403.613

Gibbs, R. W. Jr., and Nayak, N. P. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cogn. Psychol.* 21, 100–138. doi: 10.1016/0010-0285(89)90004-2

Gibbs, R. W. Jr., Nayak, N. P., and Cutting, C. (1989). How to kick the bucket and not decompose: analyzability and idiom processing. *J. Memory Lang.* 28, 576–593. doi: 10.1016/0749-596X(89)90014-4

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability,* Vol. 43. London: Chapman and Hall.

Hung, P. F., and Nippold, M. A. (2014). Idiom understanding in adulthood: Examining age-related differences. *Clin. Lingu. Phonet.* 28, 208–221. doi: 10.3109/02699206.2013.850117

Jackendoff, R. (1995). "The boundaries of the lexicon," in *Idioms: Structural and Psychological Perspectives,* eds M. Everaert, E. J. Van der Linden, R. Schreuder, and R. Schreuder (Hillsdale, NJ: Lawrence Erlbaum Associates), 133–166.

Jackendoff, R. (1997). Twistin'the night away. *Language* 73, 534–559.

Janssen, N., and Barber, H. A. (2012). Phrase frequency effects in language production. *PLoS ONE* 7:e33202. doi: 10.1371/journal.pone.0033202

Kuiper, K., Columbus, G., and Schmitt, N. (2009). "The acquisition of phrasal vocabulary," in *Language Acquisition*, eds Foster-Cohen and Susan (London: Palgrave Macmillan), 216–240. doi: 10.1057/9780230240780_10

Levorato, M. C., and Cacciari, C. (1992). Children's comprehension and production of idioms: the role of context and familiarity. *J. Child Lang.* 19, 415–433. doi: 10.1017/S0305000900011478

Li, D., Zhang, Y., and Wang, X. (2016). Descriptive norms for 350 Chinese idioms with seven syntactic structures. *Behav. Res. Methods* 48, 1678–1693. doi: 10.3758/s13428-015-0692-y

Libben, M. R., and Titone, D. A. (2008). The multidetermined nature of idiom processing. *Mem. Cogn.* 36, 1103–1121. doi: 10.3758/MC.36.6.1103

Moon, R. (1998). *Fixed Expressions and Idioms in English : A Corpus-Based Approach*. *Oxford Studies in Lexicography and Lexicology.* Oxford: Clarendon.

Nippold, M. A., and Duthie, J. K. (2003). Mental imagery and idiom comprehension: a comparison of school-age children and adults. *J. Speech Lang. Hear. Res.* 46, 788–799. doi: 10.1044/1092-4388(2003/062)

Nippold, M. A., and Martin, S. T. (1989). Idiom interpretation in isolation versus context: a developmental study with adolescents. *J. Speech Lang. Hear. Res.* 32, 59–66. doi: 10.1044/jshr.3201.59

Nippold, M. A., and Rudzinski, M. (1993). Familiarity and transparency in idiom explanation: a developmental study of children and adolescents. *J. Speech Lang. Hear. Res.* 36, 728–737. doi: 10.1044/jshr.3604.728

Nippold, M. A., and Taylor, C. L. (1995). Idiom understanding in youth: Further examination of familiarity and transparency. *J. Speech Lang. Hear. Res.* 38, 426–433. doi: 10.1044/jshr.3802.426

Nordmann, E., and Jambazova, A. A. (2017). Normative data for idiomatic expressions. *Behav. Res. Methods* 49, 198–215. doi: 10.3758/s13428-016-0705-5

Pawley, A., and Syder, F. H. (1983). "Two puzzles for linguistic theory: Nativelike selection and nativelike fluency," in *Language and Communication*, eds J. C. Richards and R. W. Schmidt (New York, NY: Longman Inc,), 191–226.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available online at: https://www.R-project.org/

Rommers, J., Dijkstra, T., and Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: evidence from idiom comprehension. *J. Cogn. Neurosci.* 25, 762–776. doi: 10.1162/jocn_a_00337

Sprenger, S., and van Rijn, H. (2013). It's time to do the math: Computation and retrieval in phrase production. *Mental Lexicon* 8, 1–25. doi: 10.1075/ml.8.1.01spr

Sprenger, S. A., Levelt, W. J., and Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *J. Memory Lang.* 54, 161–184. doi: 10.1016/j.jml.2005.11.001

Swinney, D. A., and Cutler, A. (1979). The access and processing of idiomatic expressions. *J. Verbal Learn. Verbal Behav.* 18, 523–534. doi: 10.1016/S0022-5371(79)90284-6

Tabossi, P., Arduino, L., and Fanari, R. (2011). Descriptive norms for 245 Italian idiomatic expressions. *Behav. Res. Methods* 43, 110–123. doi: 10.3758/s13428-010-0018-z

Titone, D., and Libben, M. (2014). Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: a cross-modal priming investigation. *Mental Lexicon* 9, 473–496. doi: 10.1075/ml.9.3.05tit

Titone, D., Lovseth, K., Kasparian, K., and Tiv, M. (2019). Are figurative interpretations of idioms directly retrieved, compositionally built, or both? Evidence from eye movement measures of reading. *Can. J. Exp. Psychol.* doi: 10.1037/cep0000175

Titone, D. A., and Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: familiarity, compositionality, predictability, and literality. *Metaphor. Symbol.* 9, 247–270. doi: 10.1207/s15327868ms0904_1

Titone, D. A., and Connine, C. M. (1999). On the compositional and noncompositional nature of idiomatic expressions. *J. Prag.* 31, 1655–1674.

Tremblay, A., and Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *Mental Lexicon* 6, 302–324. doi: 10.1075/ml.6.2.04tre

Van Noord, G., Bouma, G., Van Eynde, F., De Kok, D., Van der Linde, J., Schuurman, I., et al. (2013). "Large scale syntactic annotation of written Dutch: Lassy," in: *Essential Speech and Language Technology for Dutch* (Berlin; Heidelberg: Springer), 147–164. Available online at: https://www.let.rug.nl/vannoord/Lassy/

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., and Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends Hear. Sci.* 23:2331216519832483. doi: 10.1177/2331216519832483

van Rij, J., Vaci, N., Wurm, L. H., and Feldman, L., B. (in press) "Alternative quantitative methods in psycholinguistics: Implications for theory and design," in *Word Knowledge and Word Usage: a Cross-disciplinary Guide to the Mental Lexicon,* eds V. Pirrelli, I. Plag, and W. U. Dressler (Berlin: Mouton de Gruyter).

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2017). *itsadug: Interpreting Time Series and Autocorrelated Data Using Gamms*. Published on the Comprehensive R Archive Network (CRAN). Available online at: https://cran.r-project.org/web/packages/itsadug.

Westbury, C., and Titone, D. (2011). Idiom literality judgments in younger and older adults: age-related effects in resolving semantic interference. *Psychol. Aging* 26, 467–474. doi: 10.1037/a0022438

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between l1 and l2 speakers of english. *J. Phonet.* 70, 86–116. doi: 10.1016/j.wocn.2018.03.002

Wood, S. (2017). *Generalized Additive Models: An Introduction With R, Second Edition. Chapman & Hall/CRC Texts in Statistical Science*. Boca Raton, FL: CRC Press.

Wood, S. N., Pya, N., and Saefken, B. (2016). Smoothing parameter and model selection for general smooth models. *J. Am. Statist. Assoc.* 111, 1548–1575. doi: 10.1080/01621459.2016.1180986

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.